

Universidad Nacional Autónoma de México

# Ensambladores y gráficas de De Bruijn

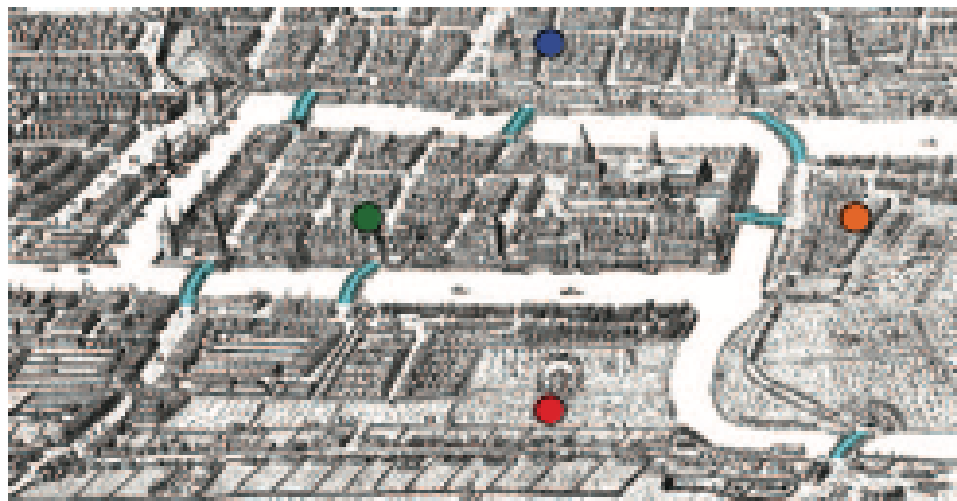
GENÓMICA COMPUTACIONAL



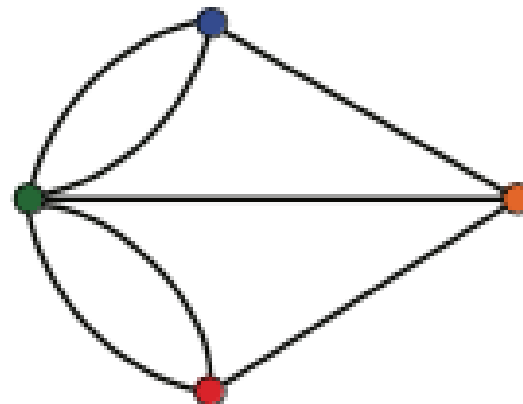
# Historia de los puentes de Königsberg

Los habitantes de esta ciudad se preguntaban si cada parte de la ciudad podía ser visitada cruzando cada uno de los siete puentes exactamente una vez y volviendo a la ubicación de partida.

a



b

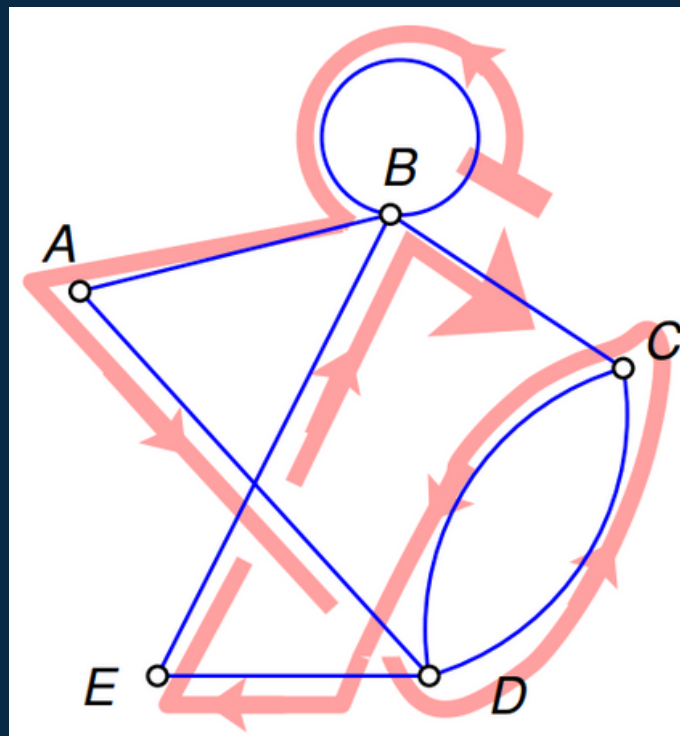


Problema de los puentes de Königsberg. (a) Un mapa de la antigua Königsberg, en el que cada área de la ciudad está etiquetada con un punto de color diferente. (b) La gráfica del puente de Königsberg, formado por la representación de cada una de las cuatro áreas terrestres como un nodo y cada uno de los siete puentes de la ciudad como un borde.

~~~~~

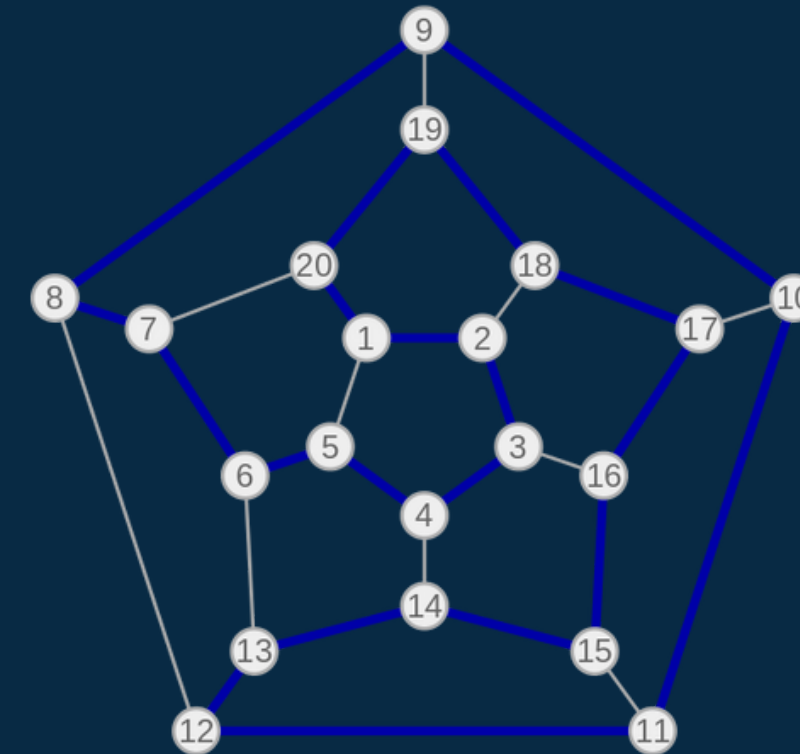
# Paseo Euleriano

Es un camino que pasa por todas las aristas de la gráfica una sola vez. Si el vértice final e inicial son el mismo, se llama Paseo Euleriano Cerrado.



# Ciclo Hamiltoniano

Este es un camino que pasa por cada uno de los nodos exactamente una vez y termina en el nodo en el que empezó.

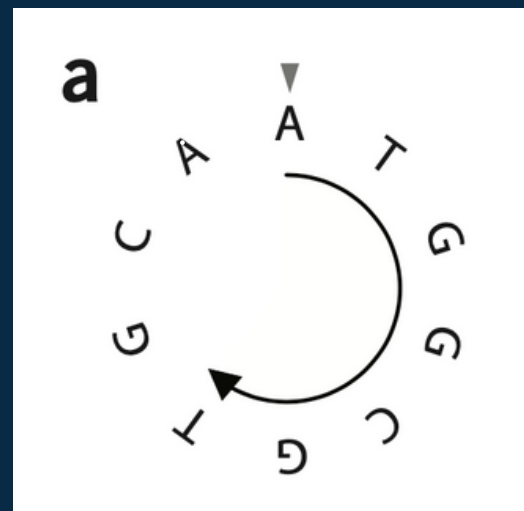


# Problemas con el ensamblaje basado en alineación

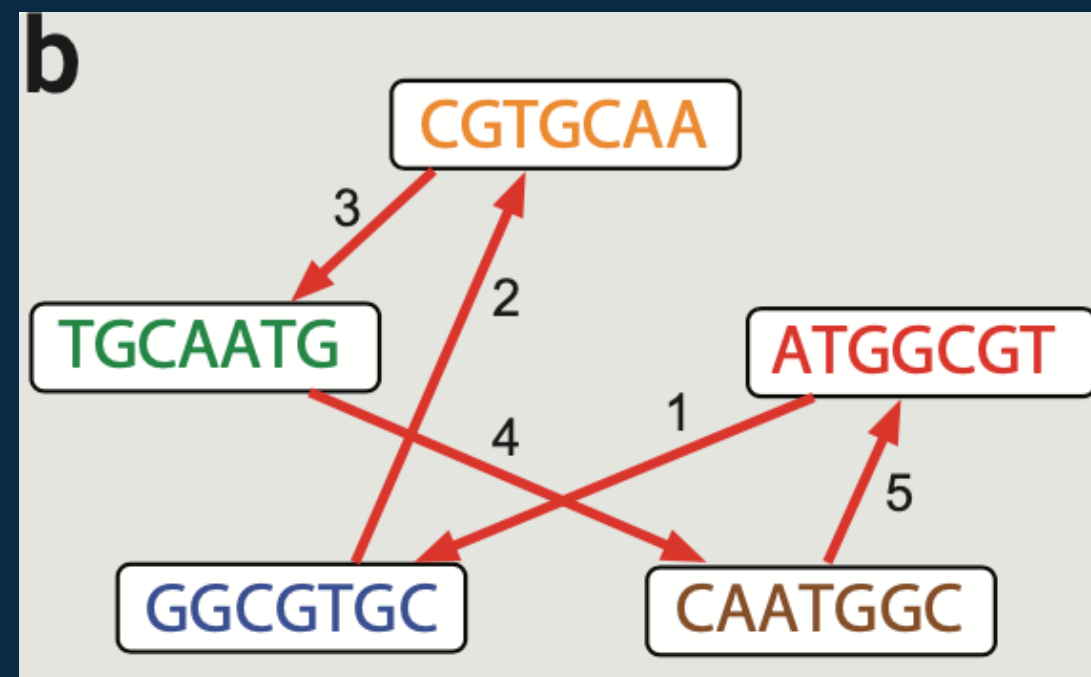
Consideremos un ejemplo sencillo de un pequeño genoma circular  
ATGGCGTGCA

Con 5 lecturas cortas :

1. CGTGCAA
2. ATGGCGT
3. CAATGGC
4. GGCGTGC
5. TGCAATG



6.



Cada lectura corresponda un vértice .  
 Dos vértices se conectan con una flecha si las lecturas  
 se superponen en al menos cinco nucleótidos.

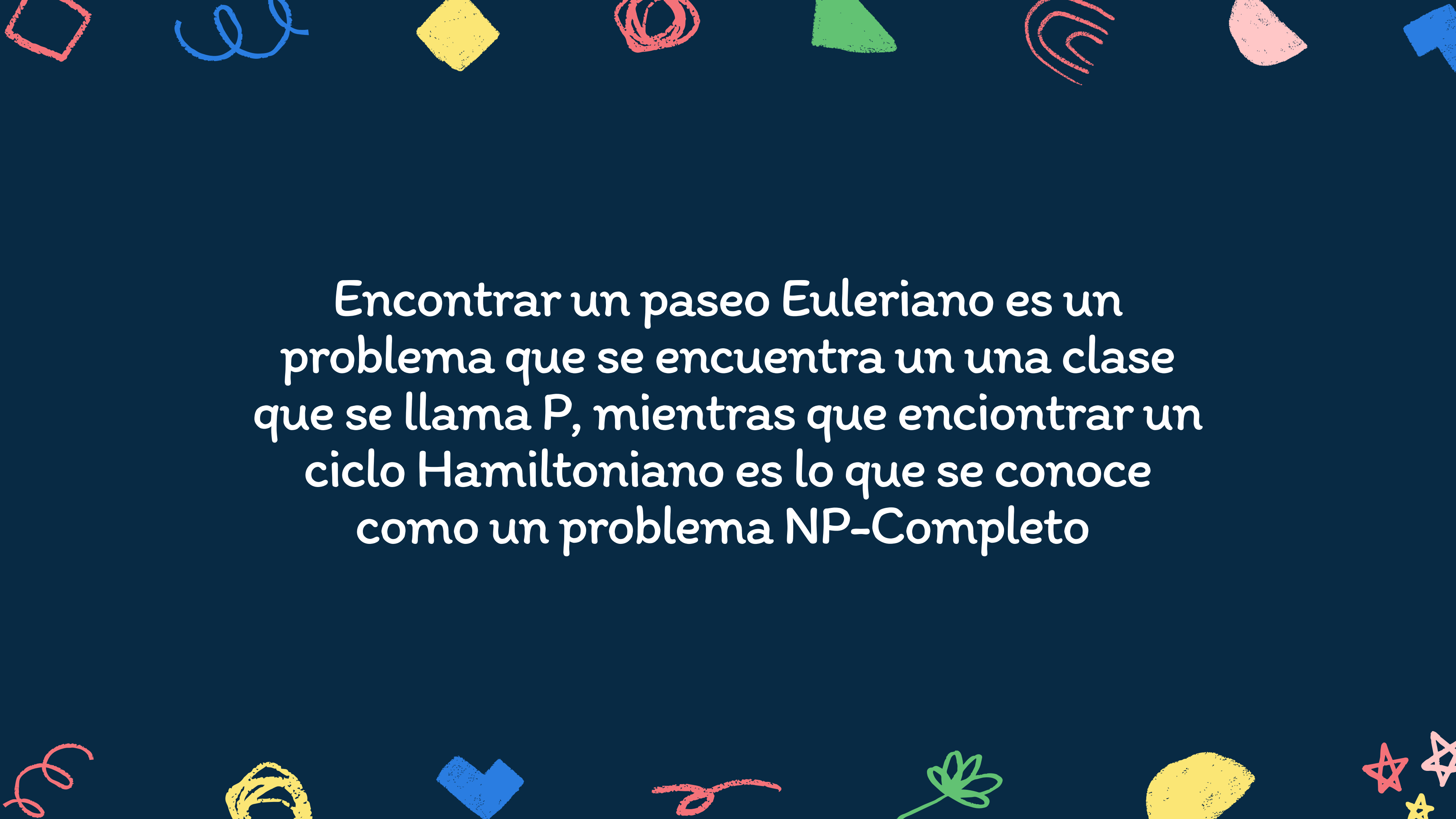


Siguiendo este camino podemos notar que se induce un  
 ciclo Hamiltoniano. El genoma circular ATGGCGTGCA, que  
 se calcula concatenando los dos primeros nucleótidos de  
 cada lectura en un ciclo hamiltoniano, contiene las cinco  
 lecturas y, por lo tanto, reconstruye el genoma original.



Este método se puede generalizar para distintos tamaños de k-mers, pero no es fácil de implementar ya que no conocemos ningún algoritmo "eficiente" para encontrar ciclos Hamiltonianos, sin embargo este método sirvió para secuenciar el genoma humano en 2001



The image features a dark blue background with a decorative border of colorful, hand-drawn shapes and patterns. The top border includes a red square, a blue swirl, a yellow diamond, a red spiral, a green triangle, a red concentric arc, a pink semi-circle, and a blue arrow. The bottom border includes a red swirl, a yellow spiral, a blue heart, a red squiggle, a green flower, a yellow blob, and several red and yellow stars.

Encontrar un paseo Euleriano es un problema que se encuentra en una clase que se llama P, mientras que encontrar un ciclo Hamiltoniano es lo que se conoce como un problema NP-Completo





# Gráficas de De Bruijn

## PASEO EULERIANO VS. CICLO HAMILTONIANO



Dado que es mucho más sencillo encontrar un paseo euleriano que un ciclo hamiltoniano, se motivó presentar el ensamblaje de fragmentos de ADN como un problema de este tipo. Para esto, usaremos gráficas de De Bruijn.

# ¿Cómo construir una gráfica de De Bruijn?



## 1. PRINCIPAL DIFERENCIA

En lugar de asignar a cada k-mer contenido en alguna lectura a un vértice, ahora asignaremos a cada k-mer a una flecha.



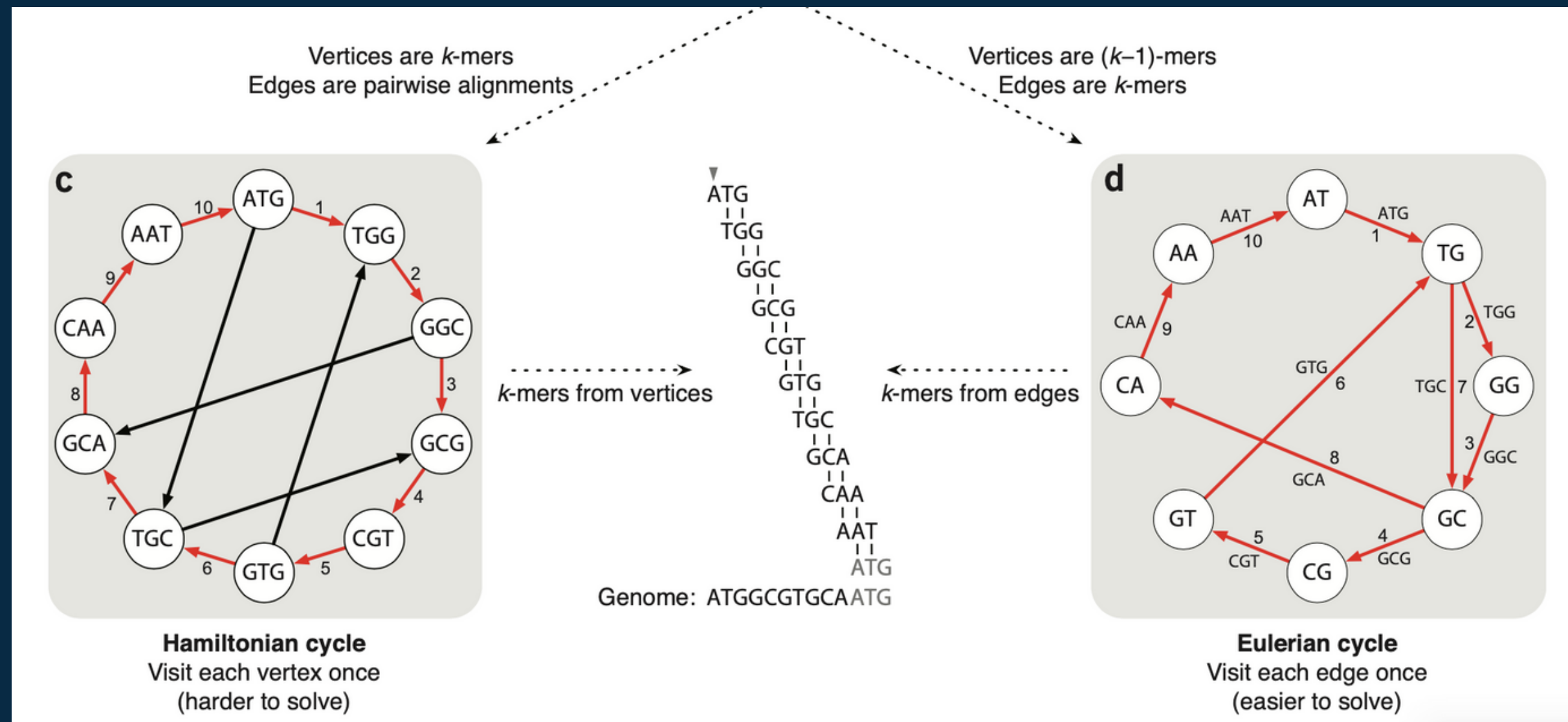
## 2. VÉRTICES

Se forma un vértice para cada prefijo o sufijo (de tamaño  $k-1$ ) distinto de un k-mer.



## 3. FLECHAS

Luego, se conecta un vértice  $x$  al vértice  $y$  con una flecha si algún k-mer (p. ej., ATG) tiene el prefijo  $x$  (p. ej., AT) y el sufijo  $y$  (p. ej., TG), y se etiqueta la flecha con este k-mer.





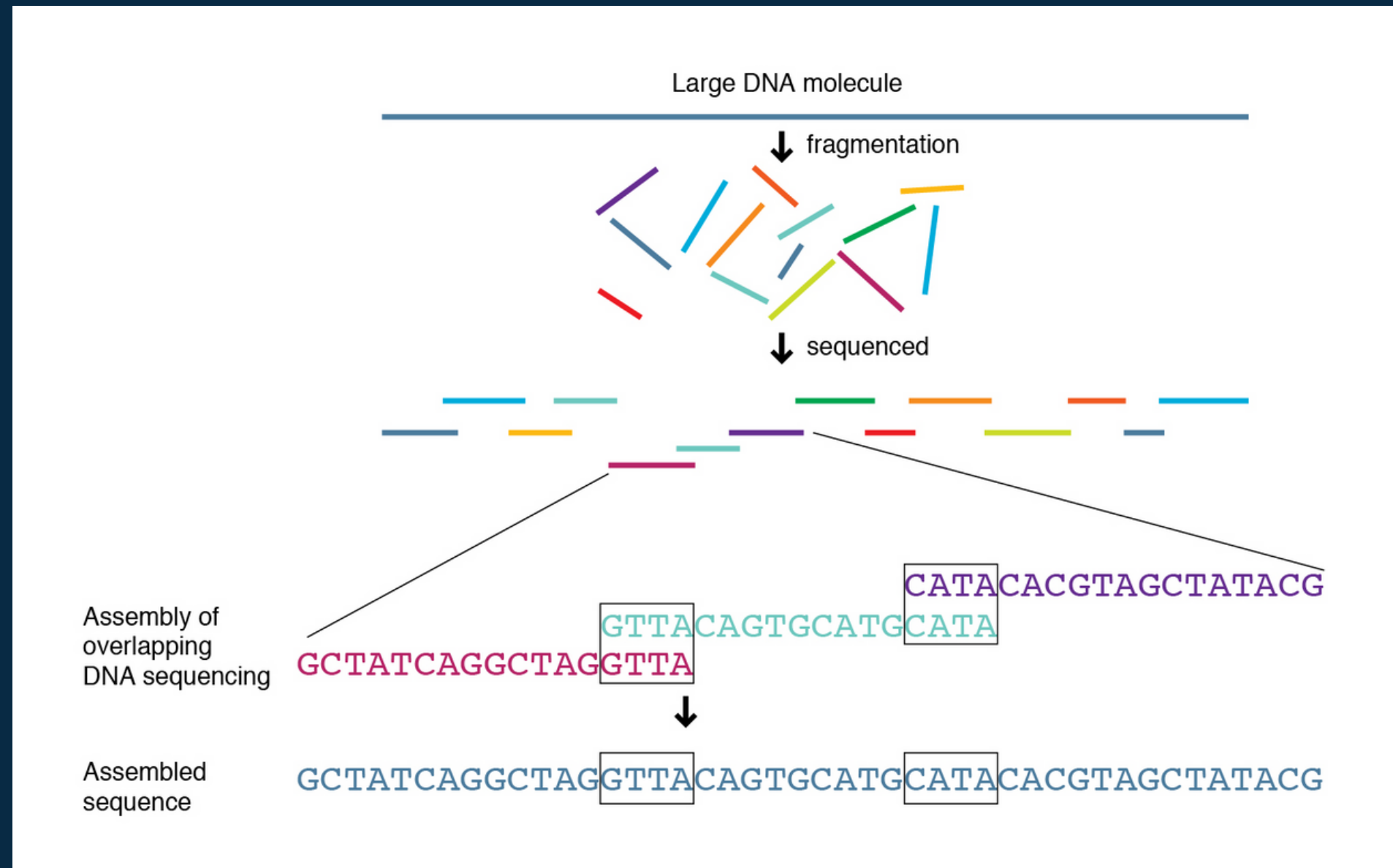
# IDBA

ITERATIVE DE BRUIJN GRAPH SHORT READ ASSEMBLER

ITERATIVE DE BRUIJN GRAPH DE NOVO ASSEMBLER FOR SEQUENCE ASSEMBLY

# De novo assembler

# Shotgun sequencing



# Tipos de ensambladores "de novo assemblers"

## Greedy algorithms

- Local optima: best overlapped short reads
- overlap-layout-consensus (Overlap graph) [OLC]

## Graph method

- String graph
- De Bruijn Graph



# Problematicas en "de novo assemblers"

False Positive  
Vertices

Branching Problems

Gap Problems





IDBA

Motivaciones

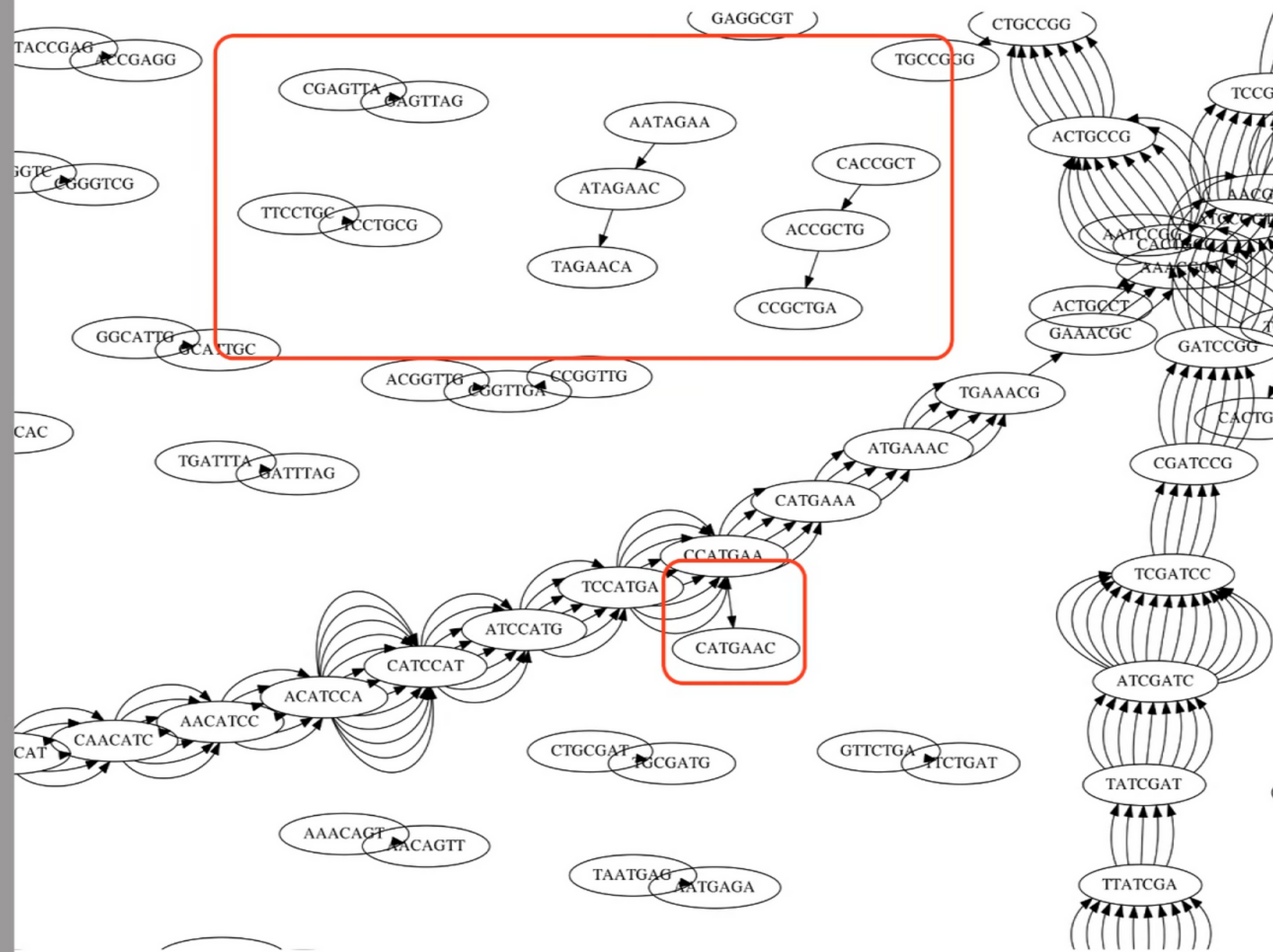
Propuestas

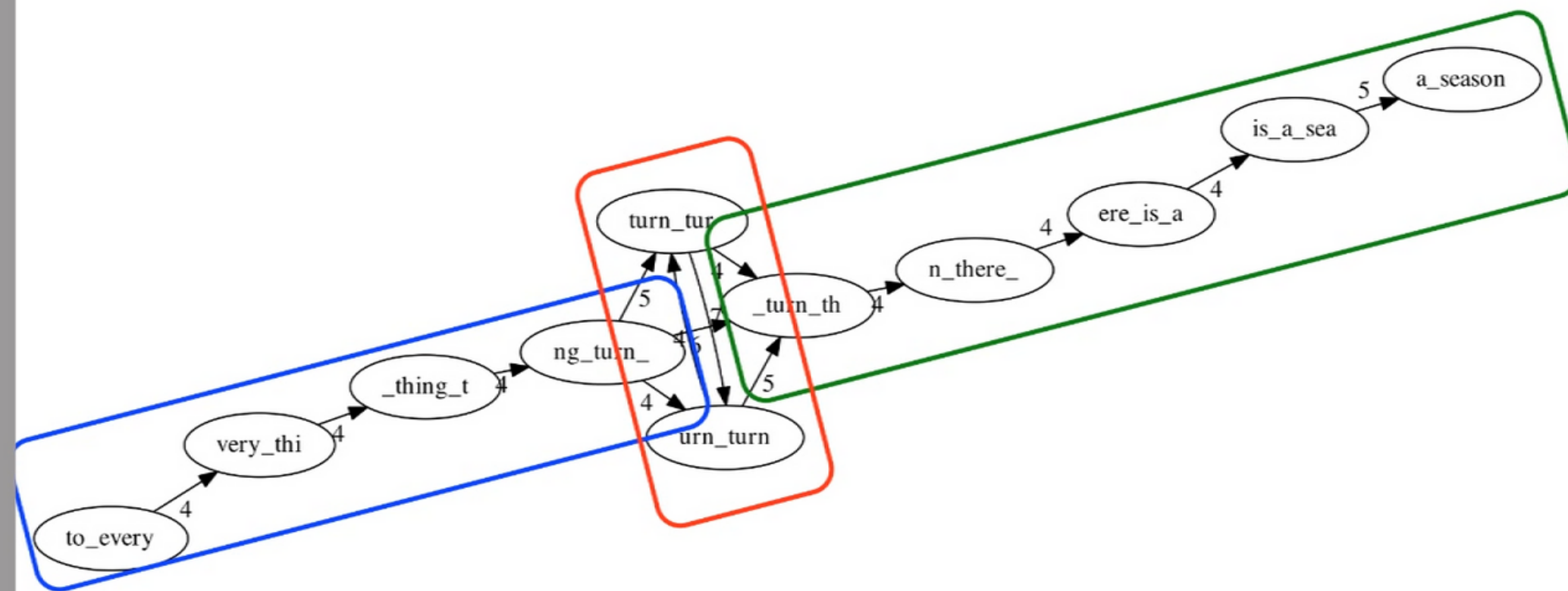
¿De qué va?

# IDBA Algorithm

## Algorithm IDBA:

```
1   $k \leftarrow k_{\min}$  ( $k_{\min} = 25$  by default)
2  Filter out  $k$ -mers appearing  $\leq m$  times
3  Construct  $H_{k_{\min}}$ 
4  Repeat
5      a) Remove dead-ends with length  $< 2k$ 
6      b) Get all potential contigs
7      c) Remove reads represented by potential contigs
8      d) Construct  $H_{k+s}$  ( $s = 1$  by default)
9      e)  $k \leftarrow k + s$ 
10 Stop if  $k \geq k_{\max}$  ( $k_{\max} = 50$  by default)
11 Remove dead-end with length shorter than  $2k_{\max}$ 
12 Merge bubbles
13 Connect potential contigs in  $H_{k_{\max}}$  using mate-pair information
14 Output all contigs
```





to\_every\_thi\_thing\_turn\_

\_turn\_there\_is\_a\_season

\_turn (repeated)



# Resultados Conclusiones IDBA



# Relevancia para la Biología



El estudio de características compartidas por diferentes genomas es fundamental para muchas áreas de la biología, como el análisis pangenómico y la genómica comparativa.





# Comparación de secuencias sin alineamiento

## \* LLAMADA DE VARIANTES Y GENOTIPIFICADO

Identificación de polimorfismos de un solo nucleótido (SNP) y pequeñas inserciones y deleciones (indels) a partir de datos de secuenciación de próxima generación.

- Predicción de la respuesta de un individuo a ciertos medicamentos
- Susceptibilidad a factores ambientales como las toxinas
- Riesgo de desarrollar enfermedades particulares.
- Rastreo de la herencia de genes de enfermedades dentro de las familias.
- Controlar la propagación de patógenos rastreando el origen de los brotes.

# Comparación de secuencias sin alineamiento

## ✿ ESTIMACIÓN DE LA ABUNDANCIA DE TRANSCRITOS

Estimar la abundancia relativa de genes que consisten en múltiples transcritos (correspondientes a diferentes isoformas) directamente del número total de lecturas mapeadas al locus del gen.

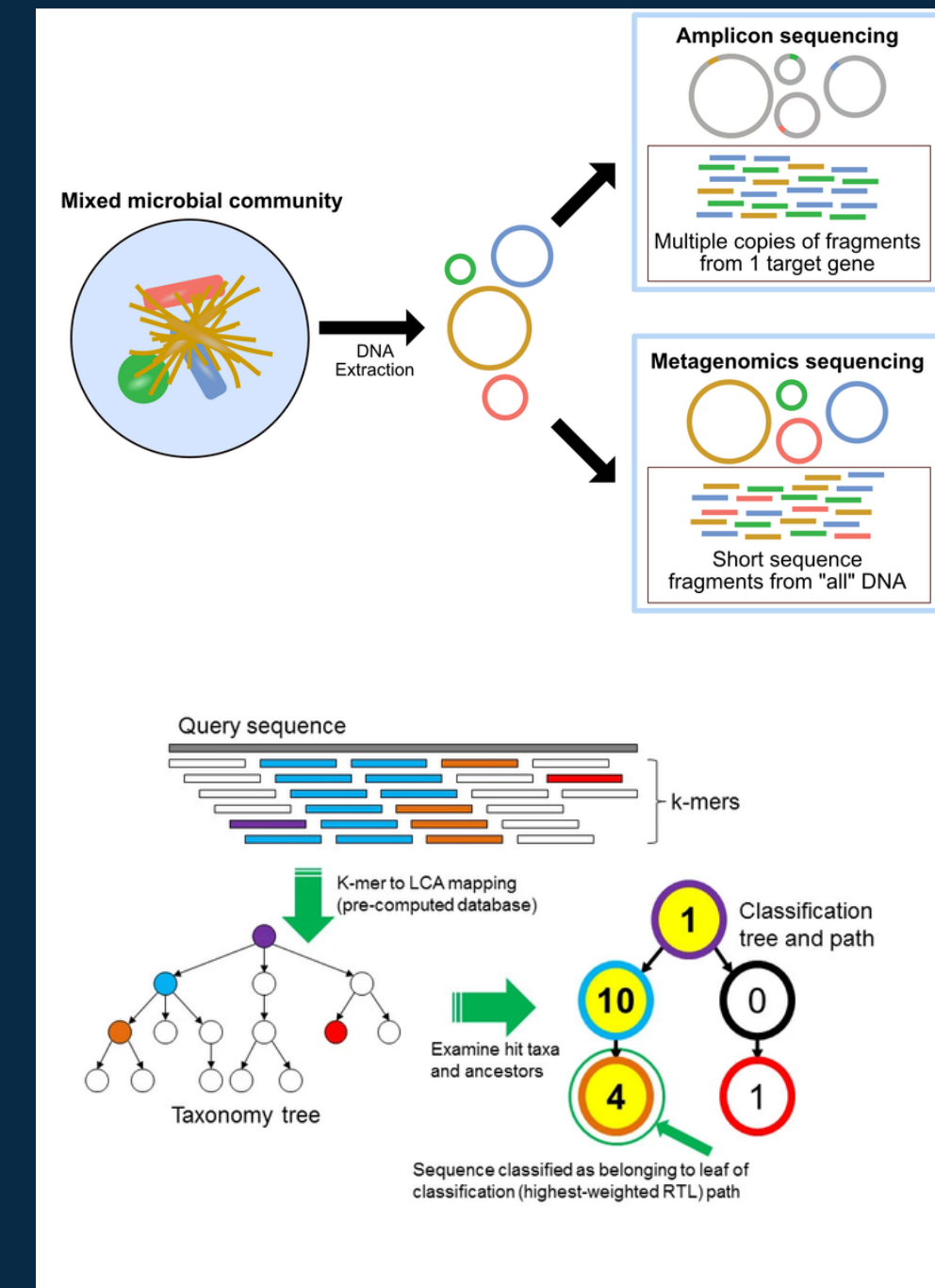
Identificación de cambios estadísticamente significativos en la abundancia relativa de transcritos al comparar diferentes experimentos.

# Comparación de secuencias sin alineamiento

## ✿ CLASIFICACIÓN METAGENÓMICA

Asignar una identidad taxonómica a cada lectura en un conjunto de datos.

Debido a que los datos metagenómicos a menudo contienen decenas de millones de lecturas, la clasificación generalmente se realiza mediante la coincidencia exacta de k-mers en lugar de usar un alineamiento de referencia, lo que sería muy lento.



# Elección de algoritmos



A medida que surgen nuevas tecnologías de secuenciación, las mejores estrategias computacionales para ensamblar genomas a partir de lecturas pueden cambiar.

- Cantidad de datos (medidos por la longitud de lectura y la cobertura)
- Calidad de los datos (incluidas las tasas de error)
- Estructura del genoma (como el número y el tamaño de las regiones repetidas y el contenido de GC).

# Idoneidad de los gráficos de de Bruijn



Las tecnologías de secuenciación de lectura corta producen un gran número de lecturas, lo que actualmente favorece el uso de gráficos de de Bruijn.

Los gráficos de de Bruijn también son adecuados para representar genomas con repeticiones, mientras que los métodos de superposición necesitan enmascarar las repeticiones que son más largas que la longitud de lectura.

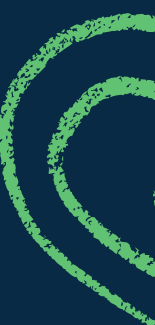
Sin embargo, si una futura tecnología de secuenciación produce lecturas de alta calidad con decenas de miles de bases, se necesitaría un número menor de lecturas y el péndulo podría volver a favorecer los enfoques de ensamblaje basados en la superposición.



# Métodos de secuenciación de alto rendimiento



| Method                                                     | Read length                                                                                                                   | Accuracy (single read not consensus)  | Reads per run                                                                                                                                       | Time per run                                                                             | Cost per 1 billion bases (in US\$) | Advantages                                                                                 | Disadvantages                                                                                                                               |
|------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|------------------------------------|--------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Single-molecule real-time sequencing (Pacific Biosciences) | 30,000 bp (N50);<br>maximum read length >100,000 bases <sup>[80][81][82]</sup>                                                | 87% raw-read accuracy <sup>[83]</sup> | 4,000,000 per Sequel 2 SMRT cell, 100–200 gigabases <sup>[80][84][85]</sup>                                                                         | 30 minutes to 20 hours <sup>[80][86]</sup>                                               | \$7.2-\$43.3                       | Fast. Detects 4mC, 5mC, 6mA. <sup>[87]</sup>                                               | Moderate throughput. Equipment can be very expensive.                                                                                       |
| Ion semiconductor (Ion Torrent sequencing)                 | up to 600 bp <sup>[88]</sup>                                                                                                  | 99.6% <sup>[89]</sup>                 | up to 80 million                                                                                                                                    | 2 hours                                                                                  | \$66.8-\$950                       | Less expensive equipment. Fast.                                                            | Homopolymer errors.                                                                                                                         |
| Pyrosequencing (454)                                       | 700 bp                                                                                                                        | 99.9%                                 | 1 million                                                                                                                                           | 24 hours                                                                                 | \$10,000                           | Long read size. Fast.                                                                      | Runs are expensive. Homopolymer errors.                                                                                                     |
| Sequencing by synthesis (Illumina)                         | MiniSeq, NextSeq: 75–300 bp;<br>MiSeq: 50–600 bp;<br>HiSeq 2500: 50–500 bp;<br>HiSeq 3/4000: 50–300 bp;<br>HiSeq X: 300 bp    | 99.9% (Phred30)                       | MiniSeq/MiSeq: 1–25 Million;<br>NextSeq: 130-00 Million;<br>HiSeq 2500: 300 million – 2 billion;<br>HiSeq 3/4000 2.5 billion;<br>HiSeq X: 3 billion | 1 to 11 days, depending upon sequencer and specified read length <sup>[90]</sup>         | \$5 to \$150                       | Potential for high sequence yield, depending upon sequencer model and desired application. | Equipment can be very expensive. Requires high concentrations of DNA.                                                                       |
| Combinatorial probe anchor synthesis (cPAS-BGI/MGI)        | BGISEQ-50: 35-50bp;<br>MGISEQ 200: 50-200bp;<br>BGISEQ-500, MGISEQ-2000: 50-300bp <sup>[91]</sup>                             | 99.9% (Phred30)                       | BGISEQ-50: 160M;<br>MGISEQ 200: 300M;<br>BGISEQ-500: 1300M per flow cell;<br>MGISEQ-2000: 375M FCS flow cell, 1500M FCL flow cell per flow cell.    | 1 to 9 days depending on instrument, read length and number of flow cells run at a time. | \$5– \$120                         |                                                                                            |                                                                                                                                             |
| Sequencing by ligation (SOLiD sequencing)                  | 50+35 or 50+50 bp                                                                                                             | 99.9%                                 | 1.2 to 1.4 billion                                                                                                                                  | 1 to 2 weeks                                                                             | \$60–130                           | Low cost per base.                                                                         | Slower than other methods. Has issues sequencing palindromic sequences. <sup>[92]</sup>                                                     |
| Nanopore Sequencing                                        | Dependent on library preparation, not the device, so user chooses read length (up to 2,272,580 bp reported <sup>[93]</sup> ). | ~92–97% single read                   | dependent on read length selected by user                                                                                                           | data streamed in real time. Choose 1 min to 48 hrs                                       | \$7–100                            | Longest individual reads. Accessible user community. Portable (Palm sized).                | Lower throughput than other machines. Single read accuracy in 90s.                                                                          |
| GenapSys Sequencing                                        | Around 150 bp single-end                                                                                                      | 99.9% (Phred30)                       | 1 to 16 million                                                                                                                                     | Around 24 hours                                                                          | \$667                              | Low-cost of instrument (\$10,000)                                                          |                                                                                                                                             |
| Chain termination (Sanger sequencing)                      | 400 to 900 bp                                                                                                                 | 99.9%                                 | N/A                                                                                                                                                 | 20 minutes to 3 hours                                                                    | \$2,400,000                        | Useful for many applications.                                                              | More expensive and impractical for larger sequencing projects. This method also requires the time-consuming step of plasmid cloning or PCR. |





# Referencias

- Compeau, P. E. C., Pevzner, P. A. y Tesler, G. (2011) "Why are de Bruijn graphs useful for genome assembly?", *Nature biotechnology*, 29(11), pp. 987-991. doi:10.1038/nbt.2023.
- Mahadik, K., Wright, C., Kulkarni, M., Bagchi, S. y Chatterji, S. (2019) "Scalable Genome Assembly through Parallel de Bruijn Graph Construction for Multiple k-mers", *Scientific Reports*. Nature Publishing Group, 9(1), p. 14882. doi:10.1038/s41598-019-51284-9.
- Nagarajan, N. y Pop, M. (2013) "Sequence assembly demystified", *Nature Reviews Genetics*. Nature Publishing Group, 14(3), pp. 157-167. doi:10.1038/nrg3367.
- Pachter, L. (2011) "Models for transcript quantification from RNA-Seq", arXiv:1104.3889 [q-bio, stat]. Disponible en: <http://arxiv.org/abs/1104.3889> (Consultado: el 2 de febrero de 2022).
- Qian, J., Marchiori, D. y Comin, M. (2018) "Fast and Sensitive Classification of Short Metagenomic Reads with SKraken", en Peixoto, N., Silveira, M., Ali, H. H., Maciel, C., y van den Broek, E. L. (eds.) *Biomedical Engineering Systems and Technologies*. Cham: Springer International Publishing (Communications in Computer and Information Science), pp. 212-226. doi:10.1007/978-3-319-94806-5\_12.
- Wood, D. E. y Salzberg, S. L. (2014) "Kraken: ultrafast metagenomic sequence classification using exact alignments", *Genome Biology*. BioMed Central, 15(3), pp. 1-12. doi:10.1186/gb-2014-15-3-r46.
- Wooley, J. C., Lin, H. S. y Biology, N. R. C. (US) C. on F. at the I. of C. and (2005) *Challenge Problems in Bioinformatics and Computational Biology from Other Reports, Catalyzing Inquiry at the Interface of Computing and Biology*. National Academies Press (US). Disponible en: <http://www.ncbi.nlm.nih.gov/books/NBK25461/> (Consultado: el 1 de febrero de 2022).
- Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. y Shen, B. (2011) "A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies", *PLOS ONE*. Public Library of Science, 6(3), p. e17915. doi:10.1371/journal.pone.0017915.